Learning Word Importance with the Neural Bag-of-Words Model
Imran Sheikh, Irina Illina, Dominique Fohr, Georges Linares

To cite this version:
Imran Sheikh, Irina Illina, Dominique Fohr, Georges Linares. Learning Word Importance with the Neural Bag-of-Words Model. ACL, Representation Learning for NLP (Repl4NLP) workshop, Aug 2016, Berlin, Germany. hal-01331720

HAL Id: hal-01331720
https://hal.archives-ouvertes.fr/hal-01331720
Submitted on 20 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

The Neural Bag-of-Words (NBOW) model performs classification with an average of the input word vectors and achieves an impressive performance. While the NBOW model learns word vectors targeted for the classification task it does not explicitly model which words are important for given task. In this paper we propose an improved NBOW model with this ability to learn task specific word importance weights. The word importance weights are learned by introducing a new weighted sum composition of the word vectors. With experiments on standard topic and sentiment classification tasks, we show that (a) our proposed model learns meaningful word importance for a given task (b) our model gives best accuracies among the BOW approaches. We also show that the learned word importance weights are comparable to tf-idf based word weights when used as features in a BOW SVM classifier.

1 Introduction

A Bag-of-Words BOW represents text (a sentence, paragraph or a document) as a vector of word features. Traditional BOW methods have used word occurrence frequency and variants of TermFrequency-InverseDocumentFrequency (tf-idf) as the word feature (Manning et al., 2008). Development in neural network and deep learning based language processing has led to the development of more powerful continuous vector representation of words (Bengio et al., 2003; Collobert and Weston, 2008; Turian et al., 2010; Mikolov et al., 2013b). It was shown that these predictive word vector representations capture syntactic and/or semantic characteristics of words and their surrounding context (Mikolov et al., 2013a; Pennington et al., 2014), and that they outperform the count based word (vector) representations (Baroni et al., 2014).

Many approaches in text classification are now driven by models built on word vectors. Earlier works proposed models which learned word vectors targeted for the classification task (Maas et al., 2011). In more recent works, text classification is performed with compositional representations learned with neural networks or by training the network specifically for text classification (Goldberg, 2015). One such network is the Neural Bag-of-Words (NBOW) model (Kalchbrenner et al., 2014; Iyyer et al., 2015). The NBOW model takes an average of the word vectors in the input text and performs classification with a logistic regression layer. Essentially the NBOW model is a fully connected feed forward network with BOW input. The averaging operation can be attributed to the absence of non-linearity at the hidden layer and the BOW inputs where words are set to 1 (or number of occurrences of that word) and 0. While the NBOW model learns word vectors targeted for the classification task, it does not explicitly model which words are important for given task.

In this paper, we propose an improved NBOW model which learns these task specific word importance weights. We replace the average with a weighted sum, where the weights applied to each word (vector) are learned during the training of the model. With experiments on sentiment and topic classification tasks, we will show that our proposed model learns meaningful word importance weights and it can perform better than the NBOW model.

The rest of the paper is organised as follows. First in Section 2 we discuss about the related works. In Section 3 we briefly introduce the NBOW model and present our proposed
model, termed as the Neural Bag-of-Weighted-Words (NBOW2) model, in Section 4. In Section 5 we give details about the experiment setup and the tasks used in our experiments. Section 6 presents a discussion on the word importance weights learned and the classification performance achieved by the proposed NBOW2 model, followed by the conclusion in Section 7.

2 Related Work

A variety of neural network architectures have been proposed for different language processing tasks. In context of text classification, fully connected feed forward networks (Le and Mikolov, 2014; Iyyer et al., 2015; Nam et al., 2014), Convolutional Neural Networks (CNN) (Kim, 2014; Johnson and Zhang, 2015; Wang et al., 2015) and also Recurrent/Recursive Neural Networks (RNN) (Socher et al., 2013; Hermann and Blunsom, 2013; Dong et al., 2014; Tai et al., 2015; Dai and Le, 2015) have been used. On one hand, the approaches based on CNN and RNN capture rich compositional information, and have been outperforming the state-of-the-art results, on the other hand they are computationally intensive and may require careful hyper-parameter selection and/or regularisation (Zhang and Wallace, 2015; Dai and Le, 2015). We focus our study to the NBOW model which gives an impressive performance in text classification, not far below the state-of-the-art CNN and RNN systems. We propose an improved NBOW model which learns these task specific word importance weights.

Word weights based on variants of word occurrence frequency or tf-idf have been commonly used and studied in literature (Manning et al., 2008; Paltoglou and Thelwall, 2010; Quan et al., 2011). Supervised weighting schemes for adjusting tf-idf for text classification have been proposed earlier (Kim and Zhang, 2014; Deng et al., 2014; Mammadov et al., 2011; Lan et al., 2006). Use of a small number of important words against all the words in the text was studied by Islam(2015) for the task of text relatedness using Latent Semantic Analysis and Google Trigram Model. Our work is in line with recent approaches of text processing with neural networks and learns word importance weights along with the word vectors.

We found that the works by Ling (2015) and Li (2014) are most related to our proposed method. Ling (2015) proposes position based weights to improve word vectors learned by the Continuous Bag-Of-Words (CBOW) model (Mikolov et al., 2013a). Li (2014) proposes Weighted Neural Network (WNN) for training RNNs which learn compositional representation of text with a parse tree. The WNN weights how much one specific node contributes to the higher-level representation. Also related are works on learning to pay attention in a sequence of input, as applied in text (Bahdanau et al., 2014) as well as speech (Chan et al., 2015), image (Xu et al., 2015) and protein sequence analysis (Sønderby et al., 2015). The model with convolutional-pooling structures presented by Gao (2014) is also shown to capture keywords in text.

3 Neural Bag-of-Words (NBOw) model

The Neural Bag-of-Words (NBOW) model (Kalchbrenner et al., 2014; Iyyer et al., 2015) is a fully connected network which maps text $X$, a sequence of words, to one of $k$ output labels. The NBOW model has $d$ dimensional word vectors for each word in the chosen task vocabulary. For the words $w \in X$, corresponding word vectors $v_w$ are looked up and a hidden vector representation $z$ is obtained as an average of the input word vectors

$$z = \frac{1}{|X|} \sum_{w \in X} v_w \quad (1)$$

The average vector $z$ is then fed to a fully connected layer to estimate probabilities for the output labels as:

$$\hat{y} = softmax(W_z z + b) \quad (2)$$

where $W_z$ is a $k \times d$ matrix, $b$ is a bias vector and $softmax(q) = \exp(q) / \sum_{j=1}^{k} \exp(q_j)$. For text classification tasks, the NBOW model is trained to minimise the categorical cross-entropy loss (Goldberg, 2015) using a stochastic gradient descent algorithm. Additional fully connected layers can be added into the NBOW model to form Deep Averaging Networks (DAN) (Iyyer et al., 2015).

4 Proposed model: Neural Bag-of-Weighted-Words (NBOW2)

While the NBOW model learns word vectors specialised for the classification task, it lacks to explicitly model and provide the information that certain words are more important than the others in
the given classification task. While tf-idf weights capture word importance weights over a given corpus and can be used at the input of the NBOW model, we are interested in letting the model learn the word importance weights which are task specific. We thus propose the NBOW2 model, with the motivation to enable the NBOW model to provide task specific word importance weights.

It is easy to realise that the NBOW model is essentially a fully connected feed forward network with BOW input vector. The averaging operation can be attributed to the absence of non-linearity at the hidden layer and the BOW inputs where words are set to 1 and 0. To learn the word importance weights, we form a weighted sum composition of the text \( X \) as follows:

\[
z = \frac{1}{|X|} \sum_{w \in X} \alpha_w v_w
\]

where \( \alpha_w \) are the scalar word importance weights for each word \( w \in X \). Learning task specific word vectors with Equation 3 ensures that words which drive the classification task are given higher importance or \( \alpha_w \) values (c.f. Figure 1).

\( \alpha_w \) are obtained by introducing a vector \( a \) in the model and are calculated as follows:

\[
\alpha_w = f(v_w \cdot a)
\]

where \( v_w \cdot a \) represents a dot product between input word vector \( v_w \) and vector \( a \), and \( f \) scales the importance weights to range \([0, 1]\). Equation 4, which makes the scalar word importance weight \( \alpha_w \) a function of the distance of the word \( w \) from \( a \) in the context space, ensures that word importance are based on contextual word similarities and not biased by frequency of occurrence in the training corpus. We refer to the vector \( a \) as the anchor vector because each word vector is compared with this vector when calculating the word importance weight.

For \( f \) common activation functions including softmax, sigmoid and also hyperbolic tangent can be used. From our experiments we found that the sigmoid function \( f(t) = (1 + e^{-t})^{-1} \) is a better choice in terms of model convergence and accuracy. However, it must be noted the softmax \( f \) could be more interesting in certain tasks because (a) the importance of a word \( (\alpha_w) \) in an input document will be dependent not only on the distance of this word from anchor vector \( a \) but also on that of the other words in the given input document (b) being a max function the softmax \( f \) will bias the composition of input document context vector \( z \) (in Equation 3) to only a handful of input words.

To summarise the as compared to the NBOW model, the NBOW2 model will include one additional anchor vector. This anchor vector is randomly initialised before training and learned along with the word vectors and other model parameters. The model training, with stochastic gradient descent, and classification (with a forward pass) both will use Equations 3 and 4, along with the output class probability estimates \( \hat{y} = \text{softmax}(W_t z + b) \) similar to the NBOW model.

5 Experiment Setup

To analyse the working and performance of our proposed NBOW2 model, we consider two common tasks: (a) binary sentiment classification on IMDB (Maas et al., 2011) and Rotten Tomatoes movie review dataset (Pang and Lee, 2005) (b) topic classification of 20 Newsgroup dataset.

5.1 Sentiment Analysis

For the IMDB task we use the original dataset\(^2\) with 25000 train and 25000 test movie reviews. For Rotten Tomatoes (RT) we obtained the v1.0 dataset\(^3\) and we do 10-fold cross-validation over the balanced binary dataset of 10,662 sentences. In both IMDB and RT tasks, model training parameters\(^4\) for NBOW2 are kept similar to those chosen for NBOW by Iyyer (2015) after cross validation. For NBOW and NBOW2 models ‘-RAND’ suffix will denote random word vector initialisation and no suffix is initialisation with publicly available 300-d GloVe vectors trained over the Common Crawl (Pennington et al., 2014)\(^5\).

5.2 20 Newsgroup Topic Classification

We use the ‘bydate’ train/test splits, cleaned and made available by Cardoso (2007)\(^6\). There are 11,293 documents in the original training set and 7,528 in the test set. For training the NBOW

\[^1\]We make available the source code at url https://drive.google.com/open?id=0BxkWDqj0McJ5R0ZPb9YVW1vWW8
\[^2\]http://ai.stanford.edu/amaas/data/sentiment/
\[^3\]https://www.cs.cornell.edu/people/pabo/movie-review-data/
\[^4\]word vector size 300, word dropout probability 0.3, L2 regularisation weight 1e-5
\[^5\]http://nlp.stanford.edu/projects/glove/
\[^6\]http://web.ist.utl.pt/acardoso/datasets/
and NBOW2 models, we randomly extract 15% of the original train set as the validation set and use remaining 85% as the final training set. Training was performed with the ADADELTA (Zeiler, 2012) gradient descent algorithm. L2 regularisation weight of 1e-5 was applied to all parameters. Further, to add robustness, we applied 75% word dropout (Iyyer et al., 2015; Dai and Le, 2015). Additionally, we use early stopping when the validation error starts to increase. Similar to the sentiment analysis experiments, '-RAND' suffix will denote random word vector initialisation and no suffix is initialisation with 300-d GloVe.

### 6 Analysis and Discussion

#### 6.1 Word importance weights learned by the NBOW2 model

We perform an analysis of the word importance weights learned by the NBOW2 model by presenting some qualitative and quantitative results.

##### 6.1.1 Visualisation of word vectors from the RT sentiment analysis task

We visually examine the word vectors learned by the NBOW and NBOW2 models. To visualise word vectors, they can be projected into a two-dimensional space using the t-Distributed Stochastic Neighbour Embedding (t-SNE) technique (van der Maaten and Hinton, 2008). Figure 1 shows the two-dimensional t-SNE visualisations of word vectors learned by the NBOW and NBOW2 models. Figure 1a shows a plot of the word vectors learned by the NBOW model and Figure 1b shows a plot of the word vectors learned by the NBOW model. Additionally in Figure 1b each word is given a colour based on the word importance assigned to it by the NBOW2 model.

From Figure 1a we can see that NBOW model tries to separate the words in the word vector space. According to the word examples labelled in Figure 1a the words appear to be grouped into regions corresponding to positive and negative sentiments of the RT movie review task. Similarly the NBOW2 model also learns to separate the words into regions of positive and negative sentiments as shown, by the same word examples, in Figure 1b. If we examine the word importance assigned by the NBOW2 model, indicated by colours in Figure 1b, it is evident that the NBOW2 model also learns to separate words based on their importance weights.

To support this statement we show additional word examples labelled in different regions in Figure 1b. For instance the words a, on, it, for, there are not so important for the RT sentiment classification task and are present together in region of lowest word importance. The words staid, inflated, softens can contribute to (negative) polarity of the reviews and hence have relatively higher importance weights (and are present together near the negative sentiment region in the word vector space).

To further verify our claim that, in comparison to the NBOW model, the NBOW2 model is able to distinguish words based on their importance we show Figure 1c. Figure 1c shows the word vectors learned by the NBOW model (same as in Figure 1a) but it depicts each word with (a colour based on) word importance weight learned by the NBOW2 model. It can be seen in Figure 1c that the NBOW model does not separate/group words based on word importance, even if we restrict only to the example words a, on, it, for, there.

##### 6.1.2 Word importance weights v/s Tf-Idf weights as classification features

In this analysis, we compare the word importance weights learned by the NBOW2 model with tf-idf weights and other word weight features proposed in the previous works. For this comparison, an SVM classifier is used for the IMDB and RT binary classification tasks. Each train/test document is represented as a sparse BOW feature vector in which each word feature is only the word weight. For NBOW2 model it is the scalar word importance weight learned by the model. We compare it with (a) classical tf-idf weights (b) credibility adjusted tf-idf (cred-tf-idf) weights proposed by Kim (2014) (c) binary cosine-normalised weights (bnc) and binary delta-smoothed-idf cosine-normalised (bΔc) weights used by Maas (2011) (d) the Naive-Bayes SVM (NBSVM) method proposed by Wang (2012). Tf-idf, bnc and bΔc word weights are task independent word weights but cred-tf-idf and NBSVM are built based on the class/task information. It should be noted that some of these methods/features have been the earlier state-of-the-art results for IMDB and RT tasks.

The classification accuracies obtained by the SVM classifiers are reported in Table 1. The tf-
Figure 1: Visualisation of word vectors learned by the NBOW and NBOW2 models in the RT task. Word vectors are reduced to 2 dimension using t-SNE technique and shown in each plot. Plot (a) represents word vectors from the NBOW model, (b) represents words from NBOW2 model, with colours indicating the word importance weights learned by the NBOW2 model, (c) represents word vectors from the NBOW model as in (a) but depicts each word with word importance weight learned by the NBOW2 model.

Table 1: Classification accuracy obtained for the IMDB and Rotten Tomatoes (RT) movie reviews sentiment classification task by training an SVM classifier on different word weights as features. (For IMDB 0.1% corresponds to 25 test documents. For RT 1% is about 10 test sentences.)

<table>
<thead>
<tr>
<th>Features for SVM Classifier</th>
<th>IMDB</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>bnc (Maas et al., 2011)</td>
<td>87.8</td>
<td>-</td>
</tr>
<tr>
<td>bΔ'c (Maas et al., 2011)</td>
<td>88.2</td>
<td>-</td>
</tr>
<tr>
<td>tf-idf-uni (Kim and Zhang, 2014)</td>
<td>86.6</td>
<td>77.1</td>
</tr>
<tr>
<td>cred-tf-idf-uni (Kim and Zhang, 2014)</td>
<td>88.8</td>
<td>77.5</td>
</tr>
<tr>
<td>NBSVM-uni (Wang and Manning, 2012)</td>
<td>88.3</td>
<td>78.1</td>
</tr>
<tr>
<td>NBOW2-RAND Word Importance Weights</td>
<td>88.2</td>
<td>76.7</td>
</tr>
<tr>
<td>NBOW2 Word Importance Weights</td>
<td>88.3</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Table 2: IMDB and Rotten Tomatoes (RT) movie reviews binary classification accuracy. First group lists BOW methods; including different initialisations of NBOW and NBOW2 (this work). The next group shows best reported results with bi-gram BOW and CNN methods, followed by LSTM RNN. Best method in each group is shown in bold. (For IMDB 0.1% corresponds to 25 test documents. For RT 1% is about 10 test sentences.)

<table>
<thead>
<tr>
<th>Model</th>
<th>IMDB</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBOW-RAND (Iyyer et al., 2015)</td>
<td>88.9</td>
<td>76.2</td>
</tr>
<tr>
<td>NBOW (Iyyer et al., 2015)</td>
<td>89.0</td>
<td>79.0</td>
</tr>
<tr>
<td>NBOW2-RAND</td>
<td>88.7</td>
<td>78.2</td>
</tr>
<tr>
<td>NBOW2</td>
<td>89.1</td>
<td>80.5</td>
</tr>
<tr>
<td>NBSVM-uni (Wang and Manning, 2012)</td>
<td>88.3</td>
<td>78.1</td>
</tr>
<tr>
<td>NBOW2-uni</td>
<td>91.2</td>
<td>79.4</td>
</tr>
<tr>
<td>CNN-MC (Kim, 2014)</td>
<td>-</td>
<td>81.1</td>
</tr>
<tr>
<td>CNN-non-static (Kim, 2014)</td>
<td>-</td>
<td>81.5</td>
</tr>
<tr>
<td>s2-bowr-CNN (Johnson and Zhang, 2015)</td>
<td>92.3</td>
<td>-</td>
</tr>
<tr>
<td>SA-LSTM (Dai and Le, 2015)</td>
<td>92.8</td>
<td>85.3</td>
</tr>
<tr>
<td>LM-LSTM (Dai and Le, 2015)</td>
<td>92.4</td>
<td>78.3</td>
</tr>
</tbody>
</table>

6.2 NBOW2 model Classification Performance

After the discussion on the word importance weights learnt by the NBOW2 model we compare the classification results obtained with our NBOW2 model. We compare the NBOW2 model classification accuracy to that obtained from the NBOW model (Iyyer et al., 2015), idf, cred-tf-idf and NBSVM methods are denoted with a ‘-uni’ suffix in Table 1 following the notation used by Kim (2014). For the SVM classifier on 25k full length documents of IMDB task, the NBOW2 model weights are as good as NBSVM and bΔ’c and better than bnc. But they do not perform as good as tf-idf weights. Whereas for the RT task with 1066 test sentences, the NBOW2 model word weights achieve accuracy closer to tf-idf variants.

BOW approaches based on Restricted Boltzmann Machines (RBM) and Support Vector Machines (SVM) and more complex approaches based on RNN, CNN. It must be noted that the CNN and RNN based approaches operate on rich word sequence information and have been shown to perform better than BOW approaches on these tasks.

Table 2 compares the classification accuracy of the NBOW2 model on IMDB and Rotten Tomatoes (RT) movie reviews binary classification tasks. Table 3 compares the classification accuracy on 20 Newsgroup topic classification. Results in Table 2 and 3 indicate that the NBOW2 model
Model | Accuracy (%)
--- | ---
NBOW-RAND | 83.2
NBOW | 83.2
NBOW2-RAND | 82.7
NBOW2 | **83.4**
RBM-MLP (Dauphin and Bengio, 2013) | 79.5
SVM + BoW (Cardoso-Cachopo, 2007) | 82.8
SA-LSTM (Dai and Le, 2015) | 84.4
LM-LSTM (Dai and Le, 2015) | **84.7**

Table 3: 20 Newsgroup topic classification accuracy. First group lists BOW methods; including different initialisations of NBOW (Iyyer et al., 2015) and NBOW2 (this work). The second group shows best reported results with LSTM RNN. Best method in each group is shown in bold. (0.2% corresponds to about 15 test set documents.)

...gives best accuracy among the BOW approaches. For IMDB and newsgroup task, the accuracy of NBOW2 model is closer to that of NBOW (not statistically significant for the 20 Newsgroup). It is also evident that for RT and newsgroup classification, the performance of NBOW2 is not far from CNN and LSTM methods.

For further analysis we also trained the NBOW2 model by simply using fixed tf-idf weights in Equation 3. This gave 87.6% and 79.4% accuracy for IMDB and RT task. Thus we can state that the word importance weights of the NBOW2 model are themselves informative.

7 Conclusion and Future Work

We proposed a novel extension to the NBOW model, which enables the model to learn task specific word importance. With experiments and analysis on sentiment and topic classification tasks, we showed that our proposed NBOW2 model learns meaningful word importance weights. We showed that the NBOW2 model gives the best accuracy among the BOW approaches and it can outperform the NBOW model. This motivates us to explore extensions to the model, including (a) class-specific anchor vectors to obtain class-specific word importance (b) document context specific word importance.

References


PA, USA. Association for Computational Linguistics.


